UNITED STATES DISTRICT COURT
DISTRICT OF MASSACHUSETTS

SCANSOFT, INC.,	)	
Plaintiff,	)	
v.	)	C.A. No. 04-10353-PBS
VOICE SIGNAL TECHNOLOGIES, INC., LAURENCE S. GILLICK, ROBERT S. ROTH, JONATHAN P. YAMRON, and MANFRED G. GRABHERR,	)	
Defendants.	)	

DECLARATION OF CHARLES C. WOOTERS IN SUPPORT OF VOICE SIGNAL TECHNOLOGIES, INC.'S OPENING CLAIM CONSTRUCTION MEMORANDUM FOR U.S. PATENT 6,594,630

I, Charles C. Wooters, on oath, depose and say as follows:

1. I am a research scientist in the field of speech recognition, and am a Senior Research Engineer at the International Computer Science Institute, Berkeley, California. I hold a Ph.D. in Speech Recognition from the University of California, Berkeley. I also hold a Master's degree in Linguistics from the University of California, Berkeley. I have published numerous articles concerning various methods and techniques in the field of speech recognition, and am a named inventor on a number of patent applications. My curriculum vitae is attached as exhibit A to this Declaration.

2. I have been retained by Voice Signal Technologies, Inc. ("Voice Signal") as a consultant in connection with this litigation. I submit this declaration in connection with Voice Signal's Opening Claim Construction Memorandum for U.S. Patent No. 6,594,630 (the '630 patent).

3. Claims 7 and 16 of the '630 patent use the term "activate" with respect to a device. In the field of speech recognition, the term "activate" means to engage the control process of a device, whatever that control mechanism is.

4. Several of the claims of the '630 patent use the term "at least one syllable in length" when describing the command word portions and pause portions of audio commands. Speech scientists have studied the length of spoken syllables in the English language and in other languages. Because of accents and idiosyncratic speech habits, the duration of one syllable can be different from speaker to speaker. For example, the one-syllable word "run" might be pronounced "run" or "ruuuuun." *See* '630 Patent, col. 3, l. 9-11. Furthermore, even the same person speaking the same phrase will not utter that phrase each time with the same syllable lengths. Certainly the same person speaking different phrases will use syllables of varying durations. Therefore, it is not possible to restrict the term "at least one syllable" to a specific measure in time. I have attached to this declaration as exhibit B a copy of a journal article describing the research work done by Steven Greenberg at the International Computer Science Institute on several characteristics of syllables. The article includes a section on the durational properties of syllables. The study shows that, over a statistically significant pool of English speech, the average syllable length is about 200 milliseconds, with a standard deviation of 100 milliseconds. In other words, as shown on the graph at figure 4 in the article, there is a range of syllable lengths around the mean, with most syllables being between 100 and 300 milliseconds in duration.

5. The ‘630 patent uses the term “spectral content.” Sound travels through the air in waves. Each wave has a frequency (measured as the distance between the repeating peaks of the waves) and an “amplitude” or energy (which can be thought of as the height of the peaks of the waves). Any given sound consists of a combination of these waves. In very simple terms, one can think, for example, of a "high" sound (a note sung by a soprano) as having relatively more energy in the high frequencies, and relatively less energy in the low frequencies. The corollary is also true. A "low sound" (a note sung by an alto) has relatively more energy in the low frequencies and relatively less energy in the high frequencies. “Spectral content” is the amount of energy at each frequency over a specified period of time. All sound has energy at different frequencies. Therefore, all sound (including command words, pauses, and background noise) has spectral content.

6. The last element of claim 7 of the ‘630 patent states that “preventing operation of the electrical device when the spectral content is dynamic.” “Dynamic” generally means changing. In the real world, the spectral content (energy at different frequencies) is always changing, even in "quiet" environments. For example, even if no one is speaking in an office environment, the background noise includes air handling systems, fluorescent lights, etc. The spectral content of even that "quiet" background noise is always changing. Therefore, spectral content is always changing. In speech recognition, we use “dynamic” to refer to changes in spectral content that are different from the change in spectral content that would be expected given the background noise.

7. If the claim were interpreted without reference to the background noise, the claimed system would never activate an electronic device, because the system would always detect some change in the spectral content during the pause portion of the audio command. It is clear,

therefore, that "dynamic" as used in the '630 patent has the meaning it has to one of ordinary skill in speech recognition, i.e. change in the spectral content that is different from the change in spectral content expected in the background noise.

8. Claim 14 of the '630 patent involves a comparison of the energy content of the command word portion to the energy of the background noise. Energy content generally correlates to what we think of as loudness. The louder the noise, the greater the total energy. Energy is a known and measurable variable. In the science of speech recognition, it is typically measured in decibels.

Sworn to under the pains and penalties of perjury this 6th day of May, 2005.

Charles C. Wooters

Exhibit A

CURRICULUM VITAE
Charles C. Wooters

Education:

Ph.D., Speech Recognition
Univ. of California Berkeley, Nov. 1993
MA, Linguistics
Univ. of California Berkeley, May 1988
BA, Linguistics
Univ. of California Berkeley, May 1986

Appointments:

Aug. 2000-Present	Senior Research Engineer International Computer Science Institute Berkeley, CA
Jul. 1999-Aug. 2000	Senior Scientist, Speech and Natural Languages GTE/BBN Technologies Columbia, MD
Apr. 1997-Jul. 1999	Senior Computer Scientist, Speech Research Division U.S. Department of Defense Ft. Meade, MD
Apr. 1995-Mar. 1997	Senior Member of Technical Staff Software Development Team Computer Motion, Inc. Goleta, CA
Dec. 1993-Mar. 1995	Senior Computer Scientist, Speech Research Division U.S. Department of Defense Ft. Meade, MD
Apr. 1993-Feb. 1994	Consultant Octel Communication Corporation Milpitas, CA
July 1988-Nov. 1993	Research Assistant International Computer Science Institute Berkeley, CA
Aug 1988-Jan. 1989	Visiting Lecturer National Tsing Hua University Hsin Chu, Taiwan, ROC
Spring 1987	Research Assistant Cognitive Studies Institute University of California Berkeley
June 1987-Aug. 1987	Research Assistant Center for Robotic Systems in Microelectronics Goleta, CA
1986-1989	Teaching Assistant, Dept. of Linguistics University of California Berkeley Berkeley, CA

Memberships:

IEEE member since 1991

Computer Skills:

C, C++, Java, Python, Perl, MySQL, Sybase, Tcl/Tk, and Matlab.

Patents:

H. Garudadri, S. Sivadas, H. Hermansky, N. Morgan, C. Wooters, A. Adami, C. Benitez, L. Burkas, S. Dupont, F. Grezl, P. Jain, S. Kajarekar, P. Motlicek. "Distributed Voice Recognition System Utilizing Multistream Network Feature Processing." Filed April 2002.

Y. Wang, D. Uecker, S. Jordan, and C. Wooters. Filed 1996. "General Purpose Distributed Operating Room Control System". Patent Pending.

Publications:

H. Shu, C. Wooters, O. Kimball, T. Colthurst, F. Richardson, S. Matsoukas, and H. Gish, 2000. "The BBN BYBLOS 2000 Conversational Mandarin LVCSR System", in NIST Speech Transcription Workshop.

M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, 1999. "Stochastic pronunciation modelling from hand-labelled phonetic corpora", Speech Communication, vol. 29, no. 2-4, Nov., pp. 209-224.

W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, 1998. "Pronunciation Modeling Using a Hand-labelled Corpus for Conversational Speech Recognition", in Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, 1997. "Pronunciation Modeling for Conversational Speech Recognition: A Status Report from WS97". In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.

J. M. Sackier, M.D., C. Wooters, L. Jacobs, M.D., A. Halverson, M.D., D. Uecker, Y. Wang. 1997. "Voice Activation of a Surgical Robotic Assistant". The American Journal of Surgery.

C. Wooters. 1995. "Duration Measurements for American English Phonemes" In "Interdisciplinary Studies on Language and Language Change: In Honor Professor William S-Y. Wang" Edited by Matthew Y. Chen and Ovid J.L. Tzeng. Pyramid Press.

D. Jurafsky, C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchman, and N. Morgan. 1995. "Using a Stochastic Context-Free Grammar as a Language Model for Speech Recognition" In Proceedings 1995 International Conference on Acoustics Speech and Signal Processing (ICASSP-95).

C. Wooters and A. Stolcke. 1994. "Multiple-Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding System." In Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP). Yokohama, Japan.

D. Jurafsky, C. Wooters, G. Tajchman, J. Segal, A. Stolcke, E. Fosler, and N. Morgan. 1994. "The Berkeley Restaurant Project" In Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP). Yokohama, Japan.

D. Jurafsky, C. Wooters, G. Tajchman, J. Segal, A. Stolcke, and N. Morgan. 1994. "Integrating Advanced Models of Syntax, Phonology, and Accent/Dialect with a Speech Recognizer", Proceedings of the AAAI Workshop on the Integration of Speech and Natural Language. Seattle WA. July 1994.

C. Wooters, 1993. "Lexical Modeling in a Speaker Independent Speech Understanding System", Berkeley, CA: University of California dissertation.

C. Wooters, D. Jurafsky, G. Tajchman, and N. Morgan, "The Berkeley Restaurant Project", pp. 119-128, Speech Research Symposium XIII, Johns Hopkins, 1993.

T. Robinson, L. Almeida, J. Boite, H. Bourlard, F. Fallside, M. Hochberg, D. Kershaw, P. Kohn, Y. Konig, N. Morgan, J. Neto, S. Renals, M. Saerens and C. Wooters, "A Neural Network Based, Speaker Independent, Large Vocabulary, Continuous Speech Recognition System: The WERNICKE Project". In Proceedings 1993 EUROSPEECH.

Y. Konig, N. Morgan, C. Wooters, V. Abrash, M. Cohen and H. Franco. 1993. "Modeling Consistency in a Speaker Independent Continuous Speech Recognition System", in Hanson J.S., Cowan J.D., and Giles C.L., editors, Advances in Neural Information Processing Systems 5, San Mateo, CA, Morgan Kaufman.

C. Wooters and N. Morgan. 1992. "Acoustic Sub-word Models in the Berkeley Restaurant Project", in Proceedings Int'l Conference on Spoken Language Processing, Banff, Alberta, Canada.

C. Wooters and N. Morgan. 1992. "Connectionist-Based Acoustic Word Models" in Proceedings IEEE Workshop on Neural Networks for Signal Processing, Copenhagen, Denmark.

H. Bourlard, N. Morgan, C. Wooters, and S. Renals. 1992. "CDNN: A Context Dependent Neural Network for Continuous Speech Recognition", in Proceedings IEEE Intl. Conference on Acoustics, Speech, and Signal Processing, San Francisco, California, March 1992.

S. Renals, N. Morgan, H. Bourlard, M. Cohen, H. Franco, C. Wooters and P. Cohen. 1991. "Connectionist Speech Recognition: Status and Prospects", ICSI Tech Report TR-91-070. December 1991.

N. Morgan, C. Wooters, and H. Hermansky. 1991. "Experiments With Temporal Resolution For Continuous Speech Recognition With Multi-Layer Perceptrons", in Proceedings IEEE Workshop on Neural Networks for Signal Processing, B.H. Juang, S.Y. Kung, Candace A. Kamn (eds.) 1991.

H. Bourlard, N. Morgan, and C. Wooters. 1991. "Connectionist Approaches to the Use of Markov Models for Speech Recognition", in Advances in Neural Information Processing Systems 3, D.S. Touretzky and R. Lippman (eds.), San Mateo, CA: Morgan Kaufman, 1991.

N. Morgan, H. Hermansky, H. Bourlard, P. Kohn and C. Wooters. 1991. "Phonetically-based Speaker Independent Continuous Speech Recognition Using PLP Analysis with Multilayer Perceptrons", in Proceedings IEEE Intl. Conference on Acoustics, Speech, Signal Processing, Toronto, Canada, 1991

N. Morgan, H. Bourlard, C. Wooters, P. Kohn, and M. Cohen, "Phonetic Context in Hybrid HMM/MLP Continuous Speech Recognition", Proceedings of Eurospeech 1991, pp. 109-112, Genova, Italy

N. Morgan, C. Wooters, H. Bourlard and M. Cohen. 1990. "Continuous Speech Recognition on the Resource Management Database Using Connectionist Probability Estimation", in Proceedings 1990 International Conference on Spoken Language Processing, Kobe, Japan, 1990.

N. Morgan, H. Hermansky, and C. Wooters. 1990. "SPOONS '90: The Speech recOgnition frOnt eNd workShop". ICSI Tech Report TR-90-045. Sept. 1990.

C. Wooters and N. Morgan 1990. "Speech Segmentation and Labeling on the NeXT Machine", ICSI Tech Report TR-90-002, January 1990.

Z.W. Shen, C. Wooters and W. S-Y Wang. 1987. "Closure Duration in the Classification of Stops: A Statistical Analysis", In: A Tribute to Ilse Lehiste, Michigan University.

Exhibit B

ELSEVIER

Speech Communication 29 (1999) 159–176

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation

Steven Greenberg [*]

International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, USA

Abstract

Current-generation automatic speech recognition (ASR) systems model spoken discourse as a quasi-linear sequence of words and phones. Because it is unusual for every phone within a word to be pronounced in a standard ("canonical") way, ASR systems often depend on a multi-pronunciation lexicon to match an acoustic sequence with a lexical unit. Since there are, in practice, many different ways for a word to be pronounced, this standard approach adds a layer of complexity and ambiguity to the decoding process which, if simplified, could potentially improve recognition performance. Systematic analysis of pronunciation variation in a corpus of spontaneous English discourse (Switchboard) demonstrates that the variation observed is more systematic at the level of the syllable than at the phonetic-segment level. Thus, syllabic onsets are realized in canonical form far more frequently than either coda or nuclear constituents. Prosodic prominence and lexical stress also appear to play an important role in pronunciation variation. The governing mechanism is likely to involve the informational valence associated with syllabic and lexical elements, and for this reason pronunciation variation offers a potential window onto the mechanisms responsible for the production and understanding of spoken language. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Automatic speech recognition; Pronunciation variation; Spoken language; Syllables

The little things are infinitely the most important – Arthur Conan Doyle (1892)

1. Introduction

No two speakers utter the same words in precisely the same way, and it is rare for the speech of even the same individual to repeat precisely over the course of a day (or even a lifetime), despite the apparent ease with which the acoustic waveform is linguistically decoded. And as aware as the listener may be of the subtle (and not-so-subtle) acoustic variations in the signal, they rarely interfere with the ability to understand spoken language. On the contrary, such variability, whether it be a consequence of the speaker's gender, age, geographical dialect or emotional state, often provides additional information with which to shape the interpretation of the signal's linguistic message. This seeming paradox, of semantic precision and informational complexity transmitted via an inherently ambiguous and variable acoustic source, is a central property of spontaneous speech, one that offers potentially keen insights into the mechanisms underlying pronunciation variation, as well as into the processes germane to the organization,

[*] Tel.: +1-510-642-4274 ext. 197; fax: +1-510-643-7684.
E-mail address: steveng@icsi.berkeley.edu (S. Greenberg)

0167-6393/99/$ - see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(99)00050-3

representation and perception of spoken language in general.

The variation of spoken language pronunciation has traditionally received relatively scant attention from the linguistic community except within the context of regional dialects (e.g., Kenyon and Knott, 1953) or sociological factors (e.g., Bernstein, 1974; Labov, 1972). Other sources of pronunciation variation have often been attributed to such factors as individual speaking style (e.g., Kohler, 1995), "economy of effort" (Jespersen, 1922) and such production factors as coarticulation (e.g., Coleman, 1992), speaking rate and reduction (e.g., Kohler, 1995; Lindblom, 1963), with relatively little effort devoted to delineating the *specific* parameters underlying its generation or linguistic expression (but, cf. Levelt, 1989; Lindblom, 1990, van Son et al., 1998). This dearth of detailed descriptive data is the likely consequence of relatively little spontaneous speech material being widely available for analysis and interpretation.

The introduction of large-vocabulary, speaker-independent speech recognition systems has resulted in the development of large-scale corpora of spontaneous speech material (e.g., Godfrey et al., 1992) and has stimulated considerable interest in pronunciation variation as a potentially important source of recognition errors (cf. Fosler-Lussier et al., 1999; McAllaster et al., 1998).

One means of dealing with variation in word pronunciation is through the creation of special-purpose lexica, incorporating some of the most commonly observed phonological variants for each word in the lexicon. Such multi-pronunciation lexica have been shown to improve the performance of such systems by a modest amount (Byrne et al., 1997; Fosler et al., 1996; Ostendorf et al., 1997; Riley et al., 1998; Weintraub et al., 1997), but not nearly to the level characteristic of human listeners.

One problem with the current multi-pronunciation approach is its emphasis on a quasi-phonemic representation for lexical elements. Individual words are represented exclusively as sequences of phones, akin to a pronouncing dictionary (e.g., Kenyon and Knott, 1953). In such lexica all elements in the phonetic sequence receive essentially equal weight relative to others. Such sequences generally reflect only a single normative ("canonical") or most common pronunciation, and it is rare that alternative lexical representations, based on organizational units above or below the phone are provided. Within such a monolithic approach lurks potentially unfortunate consequences for recognition performance when things go "wrong" (as they often do in spontaneous speech).

A recent study by McAllaster et al. (1998) demonstrates the potential significance of accurate pronunciation models for automatic speech recognition (ASR) systems. Normally, the lexicon used in an ASR system contains several of the most likely pronunciations for each word encountered by the system's decoder. Many of the pronunciations listed in the lexicon may not actually be encountered in any given recognition sequence and, conversely, some of the pronunciation variations encountered by the decoder may not be contained in the system's lexicon. The McAllaster study attempted to ascertain precisely what the impact on recognition performance would be if all of the pronunciation variants encountered by the decoder were in fact contained in the lexicon; and if they were, did it make any difference to recognition performance if the lexicon also contained pronunciation variants for these same words that did not actually occur in the data?

To address such issues phonetic sequences were "synthesized" for a corpus of telephone conversation dialogs (the Switchboard corpus, cf. (Godfrey et al., 1992)) so that the phonetic segment sequences matched the lexicon entries precisely (and conversely, all of the lexicon entries for a given word were encountered in the simulated data). In the real world such a precise match between phonetic sequence data and the system's lexicon never actually occurs. The issue of interest is whether a strict concordance between the phonetic composition of the lexicon and the actual data encountered would have a substantial impact on recognition performance.

The results of McAllaster's simulation demonstrated that word-error rate could be reduced from ca. 40% to less than 5% using this simple expedient. In other words, performance could be dramatically improved if phonetic models were *perfectly* (or at least more closely) matched to the phonetic representations contained in the lexicon. This elegant demonstration may (at least partially) explain why performance is so much better for corpora derived from read text (e.g., TIMIT, Wall Street Journal) than for those based on unscripted dialogues (cf. Bernstein et al., 1992; Weintraub et al., 1996). What clearly distinguishes these two types of spoken language material is the manner in which the speech is phonetically realized (Fosler-Lussier et al., 1999; Greenberg, 1998; Weintraub et al., 1996).

Although a "perfect" phonetic representation for each word in the lexicon may be unrealistic at present, may there not be some benefit that accrues from including at least some of the phonetically relevant variation in the AST lexicon? In order to answer this question McAllaster and colleagues included four hours of hand-labeled, phonetic transcriptions (Greenberg et al., 1996; Greenberg, 1997b) in their lexicon (cf. http://www.icsi.berkeley.edu/real/stp for a description of the specific material included in the phonetic transcription). Using this simple expedient, the error rate was reduced by about a third, indicating that even *partial* inclusion of phonetically relevant information can improve performance. However, when *all* of the phonetic variants found in the ICSI corpus were included in the lexicon, irrespective of whether they actually occurred in the test materials or not, performance actually *declined* relative to the baseline. Clearly, the contents of the lexicon needs to accurately reflect the range of phonetic variation observed; otherwise, performance is impaired.

Human listeners typically rely on many (perhaps dozens of) different linguistic tiers (e.g., articulatory-acoustic features, such as voicing, manner and place, phonetic segments, syllables, words, lexical compounds, lexical and phrasal stress, etc.) to decode the speech signal during the course of a typical conversation (Goldinger et al., 1996; Greenberg, 1997a; Levelt, 1989). Variations in the spectrum, speech envelope, fundamental frequency, segmental duration, movement of the lips and jaw, as well as detailed knowledge of the statistical and prosodic properties of spoken language are all utilized to deduce the linguistic message embedded in the acoustic signal. As of yet, ASR systems take little advantage of such extra-segmental and prosodic information in decoding the speech stream (but cf. Kompe, 1997; Niemann et al., 1997; Waibel, 1988) and it is therefore unsurprising that such features have not been systematically investigated with respect to pronunciation variation.

One means by which to rectify this representational imbalance is through systematic analysis of the *phonetic* properties of spontaneous speech in an effort to ascertain precisely how much of the variation in spoken language pronunciation can be accounted for on the basis of such narrow linguistic criteria. Such knowledge could in principle be used to delineate the extra-phonetic factors involved in the patterning of pronunciation variation and thereby improve the pronunciation models to an extent that would substantially reduce the word-error rate of current-generation ASR systems, as well as provide an empirical foundation with which to develop future-generation speech understanding systems using multiple information sources (Greenberg, 1997a).

2. The phonetic transcription of spontaneous speech

Switchboard is currently one of the primary corpora with which the reliability and accuracy of ASR systems is assessed. In contrast to such corpora as TIMIT (Zue and Seneff, 1996) or Wall Street Journal (Gauvain et al., 1994), in which a speaker reads prepared written material, Switchboard comprises informal, unscripted, telephone dialogues on a wide range of topics, incorporating a vast range of speaking styles (Godfrey et al., 1992). Moreover, the Switchboard corpus encompasses a broad range of variation in the age, gender and educational background of the speaker (comparable to that of the TIMIT and Wall Street Journal corpora).

Four hours of material from this corpus were phonetically labeled by linguistically trained, highly experienced transcribers (the interlabeler agreement ranged between 72% and 80% on the phonetic segment level – cf. (Greenberg et al., 1999) for additional detail) and made available through the Johns Hopkins' Center for Language and Speech Processing to the ASR community for developing future-generation recognition systems and for improving current methods for modeling pronunciation variation (Byrne et al., 1997, 1998; Fosler-Lussier and Morgan, 1998; Ostendorf et al., 1997; Riley and Ljolje, 1995; Riley et al., 1998; Schiel and Tillmann, 1998; Weintraub et al., 1997).

Three-quarters of the material was labeled at the phone level and segmented at syllabic boundaries. The remainder (72 minutes) was labeled and segmented at the phonetic-segment level, but also segmented at the syllabic level to insure compatibility with the remaining three hours of material. Both portions of the corpus were transcribed using a custom-designed variant (Greenberg, 1997b; Greenberg et al., 1999) of the Arpabet phonetic symbol set (Zue and Seneff, 1996). A small portion of this material was transcribed in common by all of the transcribers in order to ascertain the inter-labeler agreement. A detailed description of this project is provided in (Greenberg, 1997b; Greenberg et al., 1999). Various statistical analyses of the phonetic transcription are described in (Greenberg, 1997a; Greenberg et al., 1996). The transcription material (including audio examples) are available on the World Wide Web (http://www.icsi.berkeley.edu/real/stp).

Such analyses provide striking testimony to the ephemeral quality of the phone at the lexical level (e.g., Tables 1 and 2) by virtue of the large proportion (ca. 22%) of phonetic segments that either change their identity (i.e., substitutions) or are altogether "missing in action" (i.e., the proportion of deletions range between 9.3% and 13.6%, depending on speaking rate, with the overall proportion of deletions equal to ca. 12.5%). Occasionally, entire words are swallowed whole (<1% of the time) with only a short pausal junction marking their (perceived) location (Greenberg, 1997b). The proportion of phone substitutions ranges between 17% (slow speaking rate) and 24% (fast rate). At the level of the phone, pronunciation variation is a difficult beast to comprehend or tame. It is for this reason that the focus of the current analyses is aimed at a higher level of linguistic organization, that of the syllable.

3. Anatomy of a syllable

The syllable can be likened to a linguistic "wolf" in phonetic clothing. It is perhaps the "sheepish" nature of its outer lining that has led many to believe that it is but a mere sequence of phones and therefore can be simulated via a multiphone (i.e., tri- or quinta-phone) approach to ASR (e.g., Rabiner and Juang, 1993). What distinguishes the syllable from this phonetic exterior is its structural integrity, grounded in both the production and perception of speech and wedded to the higher tiers of linguistic organization.

The syllable is structurally divisible into three parts, the onset, nucleus and coda (the nucleus and coda, taken together, are often referred to as the "rhyme"). Although many syllables contain all three elements, a significant proportion contain only one or two. With rare exception, when a single component is present, it is the nucleus. Generally (though not always), the nucleus is vocalic, while the onset and coda are usually consonantal in form. For example, the word (and syllable) "cat", can be phonetically represented as three distinct segments, [k] [ae] [t], each associated with a specific structural element of the syllable. The [k] is the onset, followed by the nucleus [ae], with the coda, [t], bringing up the rear (cf. Crystal, 1995, p. 246). A second means by which to characterize the syllable is in terms of its consonantal-vocalic composition. Within this framework [k ae t] is classified as a CVC syllable (at least in terms of its canonical [i.e., phonological] representation).

"Real life" is often somewhat more complicated with respect to syllabic parsing of the speech stream. In particular, there is the thorny issue of "syllabification" (i.e., segmentation into syllabic entities) as well as "resyllabification" (which deals with the reassignment of a phonetic constituent (rarely more than one) from one syllable to another. These issues are of concern when comparing the phonetic transcription with the canonical forms in the lexicon. The phonetic sequences were automatically parsed into syllables using a program written by William Fisher (tsylb) (and further adapted by Dan Ellis) which instantiates algorithmically the rules for syllabification of American English described by Kahn (1980) (see

Table 1
Eighty pronunciation variants of the word "and" from the Switchboard Transcription Corpus[a]

N	Phonetic transcription					*N*	Phonetic transcription					
82	ae	n				1	ah	nx				
63	eh	n				1	ae	n	t			
45	ix	n				1	eh	d				
35	ax	n				1	ah	n	dcl	d		
34	en					1	ey	ih	n	dcl	d	
30	n					1	ae	ix	n			
20	ae	n	dcl	d		1	ae	nx	ax			
17	ih	n				1	ax	ng				
17	q	ae	n			1	ay	n				
11	ae	n	d			1	ih	ah	n	d		
7	q	eh	n			1	ae	hh				
7	ae	nx				1	ih	ng				
6	ae	ae	n			1	ix					
6	ah	n				1	ae	n	d	dcl		
5	eh	nx				1	ix	dcl	d			
4	uh	n				1	ae	eh	n			
4	ix	nx				1	hh	n				
4	q	ae	n	dcl	d	1	ix	n	t			
3	eh	n	d			1	ae	ax	n	dcl	d	
3	q	ae	nx			1	iy	eh	n			
3	eh					1	m					
2	ae	n	dcl			1	ae	ae	n	d		
2	ae					1	nx					
2	ax	m				1	q	ae	ae	n		
2	ax	n	d			1	q	ae	ae	n	dcl	d
2	ae	eh	n	dcl	d	1	q	ae	eh	n	dcl	d
2	eh	n	dcl	d		1	q	ae	ih	n		
2	ax	nx				1	aa	n				
2	q	ae	ae	n	d	1	q	ae	n	d		
2	q	ix	n			1	?	nx				
2	ix	n	dcl	d		1	q	ae	n	q		
2	ih					1	eh	n	m			
2	eh	eh	n			1	q	eh	en	dcl		
2	q	eh	nx			1	eh	ng				
2	ix	d	n			1	q	eh	n	q		
1	eh	m				1	em					
1	ax	n	dcl	d		1	q	eh	ow	m		
1	aw	n				1	q	ih	n			
1	ae	q				1	q	ix	en			
1	eh	dcl				1	er					

[a] The variants are listed in order of their frequency. The phonetic symbols are from a transcription system based on Arpabet. The segment [q] denotes a glottal stop. The symbol set and transcription methods are described in (Greenberg, 1997b, Greenberg et al., 1999).

Greenberg (1997b, 1999) for further discussion on how this syllabification algorithm was used for phonetic transcription).

If all words were spoken in canonical form, there would be little reason to prefer the syllable over some other form of multi-phone representation for ASR. However, the pattern of pronunciation variation observed in spontaneous speech is far from egalitarian. The onset portion of the syllable is generally a "survivor", maintaining its canonical identity regardless of speaking conditions, while the nucleus is a "chameleon", capable of assuming a wide range of vocalic appearances. And the coda often gets no respect, as a

Table 2
Pronunciation variability for the 100 most common words in the phonetically segmented portion of the Switchboard Transcription Corpus[a]

	Word	*N*	#Pr.	Most common pronunciation	%Tot
1	I	649	53	ay	53
2	and	521	87	ae n	16
3	the	475	76	dh ax	27
4	you	406	68	y ix	20
5	that	328	117	dh ae	11
6	a	319	28	ax	64
7	to	288	66	tcl t uw	14
8	know	249	34	n ow	56
9	of	242	44	ax v	21
10	it	240	49	ih	22
11	yeah	203	48	y ae	43
12	in	178	22	ih n	45
13	they	152	28	dh ey	60
14	do	131	30	dcl d uw	54
15	so	130	14	s ow	74
16	but	123	45	bcl b ah tcl t	12
17	is	120	24	ih z	50
18	like	119	19	l ay kcl k	46
19	have	116	22	hh ae v	54
20	was	111	24	w ah z	23
21	we	108	13	w iy	83
22	it's	101	14	ih tcl s	20
23	just	101	34	jh ix s	17
24	on	98	18	aa n	49
25	or	94	23	er	36
26	not	92	24	m aa q	24
27	think	92	23	th ih ng kcl k	32
28	for	87	19	f er	46
29	well	84	49	w eh l	23
30	what	82	40	w ah dx	14
31	about	77	46	ax bcl b aw	12
32	all	74	27	ao l	24
33	that's	74	19	dh he s	16
34	oh	74	17	ow	61
35	really	71	25	r ih l iy	45
36	one	69	8	w ah n	78
37	are	68	19	er	42
38	I'm	67	9	q aa m	26
39	right	61	21	r ay	28
40	uh	60	16	ah	41
41	them	60	18	ax m	23
42	at	59	36	ae dx	8
43	there	58	28	dh eh r	22
44	my	58	9	m ay	66
45	mean	56	10	m iy n	58
46	don't	56	21	dx ow	14
47	no	55	8	n ow	77
48	with	55	20	w ih th	35
49	if	55	18	ih f	41
50	when	54	18	w eh n	31
51	can	54	28	kcl k ae n	15

Table 2 (Continued)

	Word	*N*	#Pr.	Most common pronunciation	%Tot
52	then	51	19	dh eh n	38
53	be	50	11	bcl b iy	76
54	as	49	16	ae z	18
55	out	47	19	ae dx	22
56	kind	47	17	kcl k ax nx	21
57	because	46	31	kcl k ax z	15
58	people	45	21	pcl p iy pcl l el	44
59	go	45	5	gcl g ow	83
60	got	45	32	gcl g aa	15
61	this	44	11	dh ih s	47
62	some	43	4	s ah m	48
63	would	41	16	w ih dcl	29
64	things	41	15	th ih ng z	52
65	now	39	11	n aw	69
66	lot	39	9	l aa dx	47
67	had	39	19	hh ae dcl	24
68	how	39	11	hh aw	53
69	good	38	13	gcl g uh dcl	27
70	get	38	20	gcl g eh dx	13
71	see	37	6	s iy	80
72	from	36	10	f r ah m	28
73	he	36	7	iy	39
74	me	35	5	m iy	87
75	don't	35	21	dx ow	14
76	their	33	19	dh eh r	25
77	more	32	11	m ao r	56
78	it's	31	14	ih tcl s	20
79	that's	31	20	dh eh s	16
80	too	31	6	tcl t uw	60
81	okay	31	17	ow kcl k ey	45
82	very	30	11	v eh r iy	36
83	up	30	11	ah pcl p	34
84	been	30	11	bcl b ih n	51
85	guess	29	8	gcl g eh s	42
86	time	29	8	tcl t ay m	62
87	going	29	21	gcl g ow ih ng	13
88	into	28	20	ih n tcl t uw	14
89	those	27	12	dh ow z	42
90	here	27	11	hh iy er	25
91	did	27	13	dcl d ih dx	23
92	work	25	8	w er kcl k	66
93	other	25	14	ah dh er	26
94	an	25	12	ax n	28
95	I've	25	7	ay v	46
96	thing	24	9	th ih ng	52
97	even	24	7	iy v ix n	40
98	our	23	9	aa r	33
99	any	23	11	ix n iy	23
100	we're	23	8	w ey r	25

[a] "*N*" is the number of instances each word appears in the 72-minute corpus. "#Pr." is the number of distinct phonetic expressions for each word. "%Tot" is the percentage of the total number of pronunciations accounted for by the single most common variant. The phonetic representation is derived from a variant of the Arpabet orthography. Further details concerning both the pronunciation data and the transcription orthography may be found in (Greenberg, 1997b; Greenberg et al., 1999). From (Greenberg, 1997a).

consequence of its disposable quality. Why should this be so?

The answers are multifaceted and reside at several levels of analysis, from the acoustic-phonetic and auditory at the lower end, to the lexical, prosodic and semantic at the upper end of linguistic organization. This veil of representational tiers is governed by the requisites of information transmission and bound into a coherent, functioning whole by virtue of the syllable.

We shall first examine some of the statistical properties of the Switchboard corpus from the perspective of both the syllable and the word in order to gain some insight into the linguistic foundations of pronunciation variation. We then consider how such structure could be embedded within an ASR system and utilized for improving recognition performance. Finally, we will briefly consider how these linguistic representations might be encoded in the auditory pathway.

4. All words, great and small

Since the days of Dewey (1923) and Zipf (1945) it has been known that words differ greatly in terms of their frequency of occurrence in written language. French et al. (1930) were the first to demonstrate a comparable pattern for spoken English dialog.

A frequency analysis of the Switchboard lexicon illustrates the magnitude of this effect. The most frequent words occur far more frequently than the least (Fig. 1). The 10 most common words account for approximately 25% of all lexical occurrences in the entire corpus (ca. 140 hours of material). The 100 most frequent words account for nearly two thirds of the individual tokens in Switchboard (Fig. 2). Examination of these most frequently occurring words (Table 2) indicates that most come from the so-called "closed" set of "function" class words such as pronouns, articles, conjunctions and modal/auxiliary verbs. Many of the remainder stem from just a few basic nominal, adjectival or verbal forms. Clearly, mastery of these 100 most common words goes a long way towards facilitating comprehension of spoken discourse. The perceptual criteria for recognizing such common words are likely to be very different from those associated with their infrequent lexical counterparts. The integrity and reliability of spoken language are likely to depend on the symbiotic relationship between these two very different classes of words.

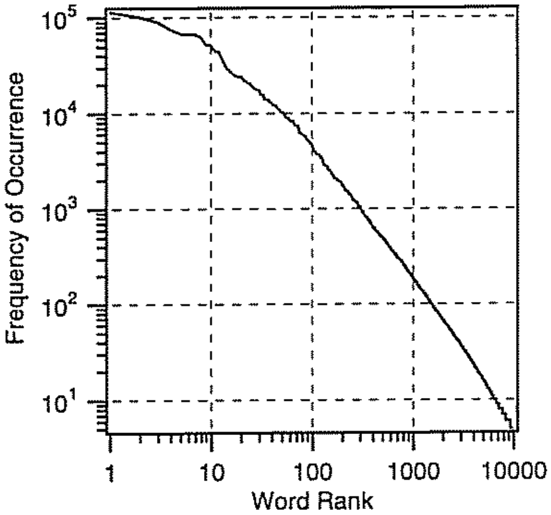

Fig. 1. The frequency of occurrence for the 10,000 most frequent words in the Switchboard corpus, organized in rank order of frequency. Total number of distinct words in the corpus is 25,923. From (Greenberg, 1997a).

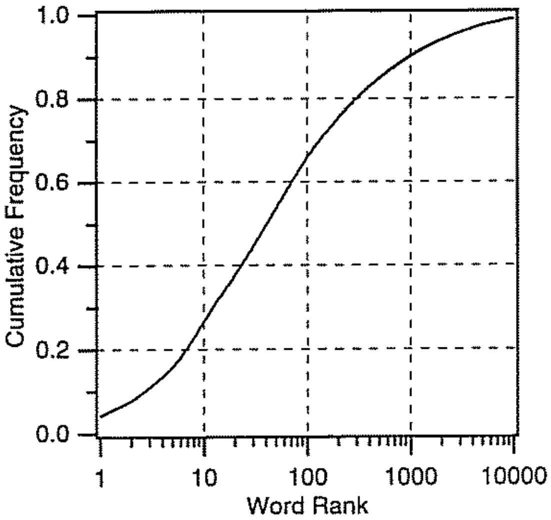

Fig. 2. Cumulative frequency of occurrence as a function of word frequency rank for the 10,000 most frequent lexical items in the Switchboard corpus. From (Greenberg, 1997a).

5. The syllabic representation of the lexicon

Although a mere list of common words does not provide sufficient data with which to interpret the speech signal, it could be used in conjunction with other knowledge sources to prune the number of likely lexical alternatives. One potentially useful representation is that of the syllable.

The 30 most common words in the Switchboard corpus are all monosyllabic (Table 2), and of the 100 most frequent lexical items only ten are not (and all of these exceptions contain just two syllables). This decided lexical preference for syllabic brevity among the most frequently occurring words is largely representative of the corpus as a whole. Although only 22% of the Switchboard lexicon is composed of monosyllabic forms, approximately 80% of the corpus tokens are just one syllable in length (Table 3). The portion of the lexicon consisting of three or more syllables (38%) is rarely exhibited in spontaneous language, accounting for less than 5% of the spoken instances (Table 3). This statistical skew towards short syllabic forms provides a potentially powerful interpretative constraint on the decoding of the speech stream.

For the 300 most frequently occurring words the cumulative statistical distribution is remarkably similar to their syllabic counterparts (Fig. 3). This identity between word and syllable is a natural consequence of the statistical relation between the two units and thus imply that segmentation and recognition of such entities effectively reduce to the same thing for this most favored portion of the lexicon.

Table 3
The proportion of words consisting of *n*-syllables for the entire Switchboard (i.e., tokens or "usage") for the *entire* (All) corpus, the portion of the corpus phonetically transcribed (STP, $n = 53{,}790$) and lexicon (i.e., type)[a]

# Syllables	Usage (%) (All)	Usage (%) (STP)	Lexicon (%)
1	81.04	78.42	22.39
2	14.30	16.31	39.76
3	3.50	3.72	24.26
4	0.96	0.95	9.91
5	0.18	0.23	3.21
6	0.02	0.03	0.40

[a] Comparable data from a telephone dialog corpus study performed in the 1920s (French et al., 1930) shows a virtually identical frequency pattern as a function of syllabic length for lexical items. From (Greenberg, 1997a).

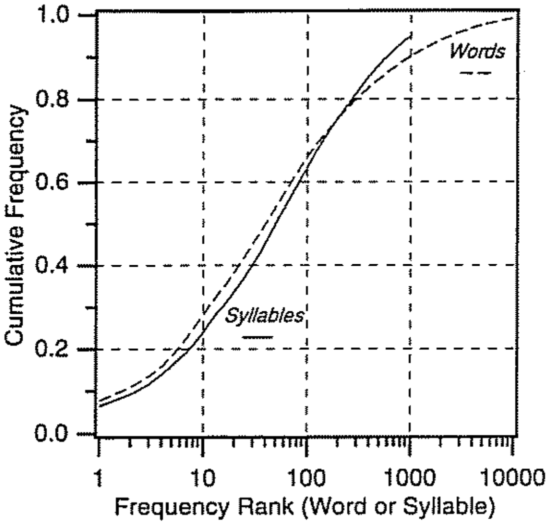

Fig. 3. The cumulative frequency of syllables in the entire Switchboard corpus as a function of syllable frequency rank compared with the cumulative frequency of occurrence for words in the same corpus. From (Greenberg, 1997a).

6. The inner life of a syllable

Many languages of the world (including Japanese and the Malayo-Polynesian family) possess a relatively transparent ("simple") syllable structure consisting of just several canonical forms. Most of the syllables in such languages contain just two phonetic segments, typically a consonantal onset followed by a vocalic nucleus (CV). The remaining syllabic forms are generally of the V (nucleus) or VC (nucleus + coda) variety. Such "syllable-timed" languages tend to exhibit an "agglutinative" (a sequence of morphemic elements obligatorily bound to the lexical root) or "isolating" (morphemic elements arranged in a relatively flexible order with respect to the word root) (cf. Lyovin, 1997) grammatical morphology and are thought to possess a relatively even tempo (but cf. Arai and Greenberg, 1997).

In contrast, English and German (as well as many other Indo-European languages) possess a more highly heterogeneous syllable structure by virtue of incorporating "complex" patterns into their syllabic repertoire. In such forms, the onset and/or coda constituents often contain two or more consonants, resulting in thousands of distinct syllabic entities (a consequence of the combinatorial potential of consonantal sequences) and tend to exhibit either an inflectional or synthetic (but rarely an agglutinative) morphology (though Lyovin (1997) suggests that English possesses certain properties more typical of an isolating language than this typology would allow). Such languages tend to informationally highlight (i.e., prosodic "prominence" or phrasal "stress") a certain proportion of syllables via selective lengthening of segmental duration, resulting in a somewhat greater variability of syllable length than observed among syllable-timed languages (cf. (Arai and Greenberg, 1997) for further discussion).

A salient property shared in common by stress- and syllable-timed languages (as exemplified by English and Japanese, respectively) is the preference for CV syllabic forms in *spontaneous* speech. Nearly half of the forms in English, and over 70% of the syllables in Japanese (Arai and Greenberg, 1997) are of this variety. There is also a substantial proportion of CVC syllables in the spontaneous speech of both languages (Table 4).

Table 4
The relative frequency of occurrence for various syllable types in both the lexicon and spoken usage of the Switchboard corpus[a]

Syllable type	Lexicon (%)	Corpus (%)	Phn. Tr (%)
"Simple"			
CV	36.2	34.0	46.30
CVC	28.8	31.6	26.36
VC	5.3	11.7	5.83
V	4.8	6.3	9.26
Subtotal	75.1	83.6	87.75
"Complex"			
CVCC	7.3	6.3	3.17
VCC	0.5	4.3	0.56
CCV	7.4	2.6	4.22
CCVC	5.0	2.2	2.42
CCVCC	2.2	0.6	0.38
CVCCC	1.0	0.4	0.14
CCCVC	0.5	<0.1	0.10
CCCV	0.4	<0.1	0.13
CCVCCC	0.3	<0.1	0.03
CCCVCC	0.2	<0.1	0.02
VCCC	<0.1	<0.1	0.01
C	N/A	N/A	0.82
CC	N/A	N/A	0.16
CCC	N/A	N/A	0.02
Subtotal	17.5	16.5	12.18

[a] The data are derived from canonical pronunciations of dictionary sources, and are compared with the syllable structure for actual pronunciation derived from phonetic transcription (Phn. Tr.). Total number of syllables = 47,406.

7. The syllabic basis of pronunciation

The importance of the syllable as an organizational unit of spoken language becomes manifest when considering pronunciation variation. In spontaneous speech the phonetic realization often differs markedly from the canonical, phonological representation (cf. Table 4). Entire phonetic elements are often dropped (28% of the consonantal codas in the phonetically transcribed portion of the Switchboard corpus) or transformed into other phonetic segments (mostly vocalic segments, ca. 35% of the time). Such patterns of deletions and substitutions appear rather complex and somewhat arbitrary when analyzed at the level of the phonetic or phonological segment. However, this variation becomes more systematic when placed within the framework of the syllable.

Several principles of pronunciation variation can be discerned for spontaneously spoken English from analyses of the Transcription Corpus, as illustrated in Tables 4 and 5:

(1) *Syllable onsets are generally preserved.* The phonetic realization of syllabic onsets tends to approximate the canonical (i.e., be "preserved") for most lexical instances and to a far greater degree, than nuclear and coda elements in *spontaneous* speech (and to a lesser extent in read text, cf. Table 5). This preference for the canonical is particularly marked for instances of complex onsets containing two or more consonantal segments (despite the greater potential for deviation from canonical pronunciation) and is most clearly

Table 5
The frequency with which the phonetic pronunciation corresponds to the lexicon's canonical pronunciation, as a function of syllabic constituent for the phonetically transcribed portion of the Switchboard corpus as well as for the TIMIT corpus of read sentential material[a]

	Switchboard		TIMIT	
Syllable constituent	All instances	Percent canonical	All instances	Percent canonical
Onset (total)	39,214	84.7	57,868	90.0
Simple [C]	32,851	84.4	42,992	88.9
Complex [CC(C)]	6,363	89.4	14,876	93.3
Nucleus	48,993	65.3	62,118	62.2
with/without onset	35,979/13,104	69.6/53.4	50,166/11,952	64.7/51.8
with/without coda	26,258/15,101	64.4/66.4	32,598/29,520	58.2/66.6
Coda (total)	32,512	63.4	40,095	81.0
Simple [C]	20,282	64.7	25,732	81.3
Complex [CC(C)]	12,230	61.2	14,363	80.5

[a] For both corpora the onsets of syllables tend to be phonetically realized as the canonical form most of the time. There is a slightly greater probability of canonical pronunciation for onsets containing two or more consonants. The vocalic nuclei are realized in canonical form far less frequently than syllabic onsets. When the syllable lacks an onset constituent the probability of canonical realization for the nucleus is significantly reduced. However, the absence of a coda element has relatively little impact on the phonetic realization of the nucleus. The primary difference in the pattern of canonical realization between read and spontaneous speech appears localized to the coda constituent. In TIMIT the coda is canonically realized nearly as often as the onset. In contrast, the coda is canonically realized significantly less frequently in Switchboard. From (Fosler-Lussier et al., 1999).

observed in the absence of a (canonical) syllabic coda, as in the case of CCV and CCCV syllabic forms (Table 4). For example, the proportion of CCV forms in the canonical lexicon is 2.6%, but rises to more than one and half times this quantity (4.2%) in terms of their phonetic realization, consistent with hypothesis that few constituents of complex onsets are deleted in spontaneous discourse (Table 4).

(2) *Coda elements are often dispensed with.* The coda element is often deleted or transformed into a segment that is phonetically homo-organic with that of the following syllable's onset (i.e., it is assimilated). The proportion of syllables classified as of the canonical CVC (31.6%) variety drops by 20% when classified on the basis of phonetic pronunciation (26.4%), reflecting coda deletion (Table 4). An even greater decline (50–80%) in the realization of the canonical coda element is observed for the VC, VCC and CVCC syllabic forms (ibid). Fully 28% of the canonical consonant codas are deleted in the transcription portion of the Switchboard corpus.

This reduction in the proportion of syllables that are phonetically realized in their canonically complex coda form is accompanied by a corresponding increase in the number of syllables that are realized without any coda at all. The proportion of CV syllables in the phonetically transcribed corpus is 46.3% despite the fact that only 34% of the canonical (i.e., phonological) forms are of this variety. And the proportion of V syllables (containing solely a nucleus) is 9.3%, even though but 6.3% of the canonical instances are of this form. The complexity of the coda, therefore, has relatively little impact on the likelihood of canonical pronunciation (in contrast to the presence of the onset constituent, cf. Table 5).

(3) *The nucleus often deviates from the canonical.* The nucleus is the syllable's "bedrock", forming its core and is virtually always vocalic in nature (only 1% of the syllables in the Switchboard Transcription Corpus lack a vocalic nucleus, cf. Table 4). Thus, any deviation from the canonical is likely to preserve the vocalic form of the nucleus, and therefore such departures are likely to be substitutions (in contrast to those of the coda, which tend to be deletions, as described above).

(4) *The likelihood of canonical expression percolates through the syllable.* The probability of

canonical pronunciation for a given constituent is influenced, to a certain degree, by the pronunciation of the other constituents in the syllable, particularly the onset (cf. Table 5). Thus, the probability of the nucleus being pronounced canonically is higher when the onset is also articulated in the standard manner (ibid). This is also the case for the coda. Furthermore, the coda is more likely to be pronounced in canonical fashion if the nucleus is as well and vice versa (cf. Greenberg, 1998, Tables 7 and 8). Such a pattern of pronunciation variation implies that the specific mechanism responsible for crafting pronunciation looks beyond the individual constituent and almost surely reflects control at the syllabic, lexical or even phrasal level. The factors potentially governing this syllabic linkage are discussed in Sections 8 and 9.

(5) *Syllabic entities are not always clearly demarcated in terms of phonetic constituents.* Such occurrences of ambisyllabicity occur in 5.9% of the syllables and refer to the situation where a phonetic constituent straddles two syllabic entities. In virtually all instances, the element shared in common is consonantal in nature and typically forms the coda of the initial syllable. Most of these instances of ambisyllabicity result in resyllabification, where the coda of the initial syllable also becomes the onset constituent of the following syllable (e.g., "for eight" > [faor] [r ey t]). However, there are a certain proportion of instances in which the coda completely detaches from the initial syllable and becomes exclusively associated with the onset of the following syllable. For this reason, the proportion of resyllabified syllables (8.2%) substantially exceeds the number of ambisyllabic forms alone.

(6) *The linguistic factors governing pronunciation variation are likely to reflect exceedingly high-level processing.* Specific patterns of pronunciation are likely to reflect the information valence of the utterance and to be indicative of the speaker's projection of the listener's internal knowledge model. This patterning is observed in the relation between lexical frequency, speaking rate, prosodic prominence and the probability of pronunciation variation (as discussed in Sections 8 and 9).

8. The role of prosodic stress in pronunciation variation

Prosody appears to have a systematic effect on pronunciation variation. In terms of the current study it manifests its influence in terms of the heightened probability of canonical pronunciation across the syllable. When a syllable is "accented" (i.e., prosodically "prominent") there is a greater tendency for the nucleus and coda constituents to be canonically pronounced than for unaccented syllables (cf. Lehiste, 1996).

8.1. Durational properties of syllables

One of the acoustic parameters most closely associated with prominence is syllable duration. Accented syllables tend to be longer in duration for several reasons. First, the nuclei of such syllables tend to be longer (Silipo and Greenberg, 1999). Second, there is a greater likelihood that all (or most) of the canonical phonological constituents will be realized (i.e., few, if any, phonetic deletions will occur for any given syllable). And finally, many of the words that tend to be accented are phonologically more complex (i.e., are composed of complex syllabic onsets and codas) as a result of their relatively higher entropy (this relation is a corollary of Zipf's "law", associating the number of phonological constituents within a word with the word's frequency of occurrence (Zipf, 1945) – most such "complex" phonological syllables come from the right-hand branch of the lexical distribution shown in Fig. 1).

The durational properties of syllables in English are illustrated in Fig. 4 for the four hours of phonetically transcribed material in the Switchboard corpus. Of interest is the broad dispersion of syllable lengths (mean = 200 ms, s.d. = 103 ms) and the extended tail of the right-hand portion of the distribution. Approximately 15% of the syllables are longer than 300 ms. Most of this population is likely to be prosodically accented.

As a consequence of this durational distribution, in concert with the pattern of conditional syllabic duration (Fig. 5), the probability of a shorter syllable following a longer one (and vice versa) is quite high (most words are one syllable

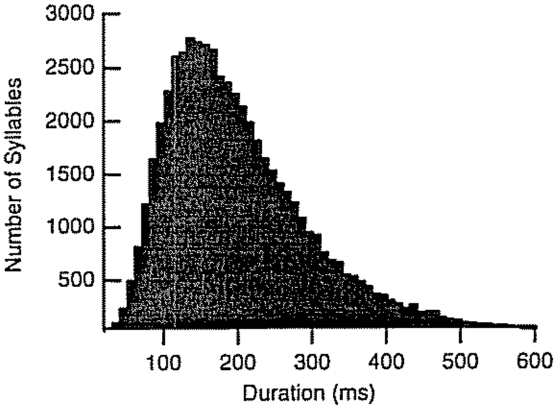

Fig. 4. Frequency distribution of syllables from a corpus of spontaneous English discourse (Switchboard). Durations are derived from manual segmentation of syllabic boundaries by phonetically trained individuals. The mean of the distribution is 200.5 ms and the standard deviations is 103 ms. $N = 56{,}400$ syllables.

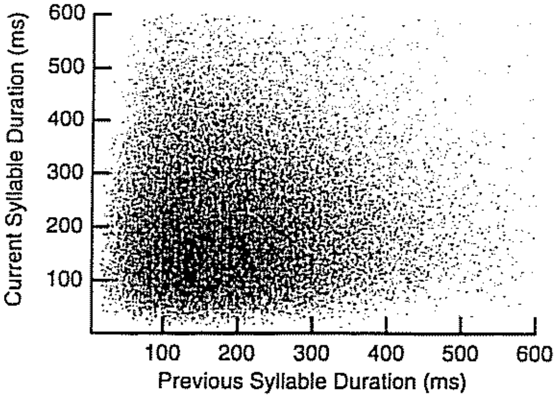

Fig. 5. Conditional dependence of syllable duration between successive syllables using the same portion of the Switchboard corpus as in Fig. 4. $N = 47{,}061$ syllable pairs. Adapted from (Arai and Greenberg, 1997).

long; therefore the conditional durations for syllables pertain mostly across words as well). Such a pattern of durational relations follows directly from the shape of the statistical distribution and does not necessarily imply any form of conscious alternation between long and short syllables. The nature of this process is illustrated in Fig. 5, which plots the data of Fig. 4 in terms of the durational relation of successive syllables. Syllables whose durations are relatively close to the mean of the distribution will tend to be followed and preceded by syllables that do not differ all that much in length. Syllables appreciably longer or shorter than the mean tend to be bracketed by syllables that are significantly different in duration. However, the conditional dependence of duration on prior and following syllable length is remarkably even and devoid of bias. The only basis for the skew in odds is the sheer weight of numbers in the distribution's core.

8.2. *The acoustic bases of prosodic stress*

The relation between prosodic stress and syllabic duration is well established in the experimental literature (cf. Lehiste, 1996) but has not previously been systematically tested on a corpus of *spontaneous* English discourse (a recent study has been conducted on spontaneous Dutch, cf. (van Kuik and Boves, 1999)). It is likely that a language's stress pattern is linked, in some measure, to the structure of its syllabic constituents as described in Section 8.1.

To ascertain duration's role in prosodic stress, several minutes of the Switchboard corpus and over two hours of a separate spontaneous corpus (consisting of ca. 1-minute monologues), OGI Stories (American English) were labeled by trained linguists in terms of accented and unaccented syllables (along with an intermediate category designating syllables that were neither fully prominent, nor completely unaccented). This material was used to train and develop an automatic algorithm to prosodically label comparable material from the same corpora. The automatic procedure is capable of correctly distinguishing accented from unaccented syllables ca. 80% of the time for the OGI Stories material and about 85% of the time for Switchboard (Silipo and Greenberg, 1999). There is ca. 84–88% agreement between labelers in distinguishing fully accented from completely unaccented syllables in the OGI Stories corpus (no interlabeler agreement was computed for the Switchboard material, as it was based on the material of only a single transcriber).

The automatic algorithm operates on three separate parameters of the acoustic signal – duration, amplitude and fundamental frequency (as well as their dyadic and triadic products). An optimization routine was developed to ascertain the combination yielding optimum discrimination between stressed and unstressed syllables. The results indicate that the product of amplitude and duration (i.e., integrated energy) of the syllabic nucleus yields the performance closest to that of the linguistic transcribers. Fundamental frequency turns out to be relatively unimportant for distinguishing between the presence and absence of prosodic prominence. These results are consistent with those reported by van Kuik and Boves (1999) for Dutch, and by Waibel (1998) for read text (in English).

8.3. The relation between prosodic stress and pronunciation variation

Prosodic stress serves to informationally highlight specific lexical and syllabic elements. Approximately one fifth of the syllables in both the OGI Stories and Switchboard material are fully accented. This proportion is likely to reflect some intrinsic division of the lexicon into informationally significant classes.

There appears to be a positive relation between canonical pronunciation and prosodic stress. The two properties often travel together, though they are not inseparable under all conditions. Syllables whose entire suite of constituents are canonically expressed are more than likely to be fully accented. However, two-thirds of the phonetic segments are pronounced in canonical fashion, three times the proportion of stressed elements, indicating that stress is only one of several factors underlying the patterning of pronunciation variation.

9. Information's role in pronunciation variation

Words of high information valence (these are typically infrequently occurring referential constituents of a nominal phrase [i.e., nouns or adjectives]) tend to be pronounced in canonical fashion, while common lexical items, particularly pronouns, conjunctions and articles, generally depart from canonical form with regularity (Fosler-Lussier and Morgan, 1998; Greenberg, 1997a). Such patterning suggests that the information valence associated with specific words and syllables may play a decisive role within an utterance (Greenberg, 1997a; van Son et al., 1998) and is therefore of potential significance for the design of ASR systems, as their lexica usually contain a single canonical pronunciation (and occasionally several alternative forms) for each lexical entry. The task of going from the acoustic signal to sequences of words for such a system would, in principle, be far simpler if each spoken instance were of the canonical form (i.e., spoken in a manner similar to that of read text). Since the probability of a word being spoken in canonical fashion increases as the speaking rate declines (Fosler-Lussier and Morgan, 1998) it is likely that the negative relationship between recognition performance and speaking rate is at least partially a consequence of this factor.

However, speaking rate, per se, may not be the sole governing factor guiding the pronunciation of a spoken utterance. The degree to which a syllable deviates from the canonical is likely to be a function of *both* speaking rate and word frequency (cf. Fosler-Lussier and Morgan, 1998). The slope of the function relating word frequency and speaking rate to the probability of canonical pronunciation is far steeper for words of high frequency than for low. In other words, as the speaking rate increases, the probability of canonical pronunciation diminishes for words of both low and high frequency. However, the impact of speaking rate is far more pronounced for frequently occurring words (with low entropy) than for rare words.

If speaking rate were the only factor involved the slope of the function would be relatively constant across word frequency. However, the slopes differ by roughly a factor of two (Fosler-Lussier and Morgan, 1998) suggesting that low-frequency words, irrespective of their rate of articulation, are more likely to be realized in canonical form than their frequently occurring counterparts.

The information valence of frequent words is also more likely to fluctuate as a function of phrasal and sentential context than less common lexical items, thus providing a potential basis for the greater variability of pronunciation under differential speaking conditions. It is likely that frequently occurring words tend to be spoken faster and in more reduced fashion because of their inherent predictability (cf. Lindblom, 1990; van Son et al., 1998). Under extreme conditions words of high frequency (and hence predictability) may be entirely deleted from the utterance, but without the listener's conscious awareness. This occurs for less than 1% of the syllables and words in Switchboard (Greenberg, 1997b). In contrast, ca. 22% of the canonical phonetic segments are not articulated. Most of these "deleted" segments are in coda position.

10. The auditory basis of pronunciation variation

Why are the onsets of syllables relatively well preserved (85%) while the codas (65%) and nuclei (63%) are so highly variable in pronunciation of spontaneous speech? Most accounts of pronunciation variation cite biomechanical constraints imposed by the vocal apparatus (cf. Levelt, 1989; Lindblom, 1990) as the controlling parameter. However, such production-based accounts do not explain why a mechanical system capable of such versatile behavior under a wide range of speaking conditions can also tailor its performance at will to deviate in systematic fashion when circumstances dictate. Might articulation serve as the handmaiden of higher linguistic function, guided in its behavior by the informational demands of the regime?

The auditory system is particularly sensitive and responsive to the beginnings of sounds, be they speech, music or noise (Greenberg, 1997c). Our sense of hearing evolved under considerable selection pressure to detect and decode constituents of the acoustic signal possessing potential biological significance. Onsets, by their very nature, are typically more informative than medial or terminal elements, serving both to alert, as well as to segment the incoming acoustic stream. They are also important for combatting the potentially deleterious effects of acoustic interference and reverberation (ibid). For these reasons the majority of auditory neurons, from periphery to cortex are most highly responsive to the initial portion of a signal. Complex, multi-level chains of neural adaptation and inhibition reinforce and enhance this bias towards the onset of events (Greenberg, 1997c; van Wieringen, 1995).

The syllable can be thought of as the structural instantiation of this auditory process, shaping the encoding of linguistic information so as to maximize its probability of reception and decoding. Over the course of a lifetime, control of pronunciation is beveled so as to take advantage of the ear's (and the brain's) predilection for onsets and to tailor the meter of the speech to the low-frequency rhythm of the auditory cortex.

11. Conclusions

Phonetic characterization of spoken language is essential for developing accurate acoustic models required for robust speech recognition under the wide range of conditions characteristic of the real world. Variability in speaking style and the acoustic background is currently the bane of data-driven, recognition systems, due to the inadequacy of lexical models derived from rigidly defined sequences of phonetic elements. In particular, current methods fail to adequately model the variance of segmental units upon which the lexical models are based.

What is required is a methodological framework for modeling pronunciation that captures the full range of phonetic phenomena characteristic of speech spoken under real-world conditions and constructs a lexical representation based on structural units that are both more "primitive" than the phonetic segment (e.g., articulatory-based features, such as manner and place) and, at the same time, capable of being integrated into a representational framework superseding that of the phone.

The variation in pronunciation observed in a corpus of spontaneous American English (Switchboard) appears more systematic when analyzed at the level of the syllable. Specifically, (a) syllable onsets are generally preserved (Table 5), (b) coda elements are often dispensed with (ibid), (c) the nucleus often deviates from the canonical (ibid) and (d) the likelihood of canonical expression percolates through the syllable (ibid). These properties of pronunciation suggest that the syllable may be a more basic organizational unit than the phone at the acoustic–phonetic level. Thus, lexical models based on syllables, supplementing the more traditional phonemic sequences commonly employed are more likely to provide a stable, robust representation of the speech signal under the wide range of acoustic and speaking conditions typical of the real world (cf. Kingsbury et al., 1998). This strategy has been successfully employed by Wu et al. (1998a,b) on a small vocabulary recognition task (OGI Numbers) to reduce the error rate by ca. 35% over the baseline (phone sequence only) system. Fosler-Lussier and coauthors have developed dynamic pronunciation models based on syllabic units for improving recognition performance on large vocabulary speech recognition (Switchboard and Broadcast News) (see e.g., Fosler-Lussier and Morgan, 1998; Fosler-Lussier et al., 1999).

Future-generation speech recognition systems are likely to benefit from incorporating syllable-level information into the decoding process, as well as from building lexical models that consist of syllabic constituents in addition to the phone sequences commonly used. In addition, much of the detailed variation in pronunciation observed in spontaneous speech is likely to be captured only at the level of acoustic-articulatory features clustered at the syllable or word level (cf. Kirchhoff, 1999; Ganapathiraju et al., 1997) using finer-grained acoustic models than are possible at the phone or syllabic level. For these reasons, modeling pronunciation variation for automatic speech recognition will require detailed quantitative characterization of spoken language on many organizational levels, ranging from acoustic features and phonetic segments, to syllables, words and phrases.

Acknowledgements

This research was supported by the National Science Foundation, the Johns Hopkins University Center for Language and Speech Processing and the US Department of Defense. The computational assistance of Takayuki Arai, Dan Ellis, Eric Fosler-Lussier, Joy Hollenback, Rosaria Silipo and Su-Lin Wu is gratefully acknowledged. I would also like to thank Colin Baker, Jeff Good and Joy Hollenback for performing the labeling of prosodic prominence. This article is based on a paper published in the Proceedings of the ESCA Tutorial and Advanced Research Workshop on Modelling pronunciation Variation for Automatic Speech Recognition (Kerkrade, Netherlands, 4–6 May 1998) as well as on material originally published in the ESCA Tutorial and Advanced Research Workshop on Robust Speech Recognition for Unknown Communication Channels (Pont-a-Mousson, France, 17–18 April 1997). Many thanks to Dr. Helmer Strik, the organizer of the Kerkrade workshop, for his many helpful suggestions for revising the original paper. The paper has also benefited considerably from the comments of three anonymous reviewers. Their time and diligence is greatly appreciated.

References

Arai, T., Greenberg, S., 1997. The temporal properties of spoken Japanese are similar to those of English. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 1011–1014.

Bernstein, B.B., 1974. Class, Codes and Control. Routledge, Kegan Paul, London.

Bernstein, J., Baldwin, G., Cohen, M., Murveit, H., Weintraub, M., 1992. Phonological studies for speech recognition. In: Proceedings of the DARPA Speech Recognition Workshop, pp. 41–48.

Byrne, W., Finke, M., Khudanpur, S., McDonnough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., 1997. Pronunciation modelling for conversational speech recognition – A status report from WS97. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 26–33.

Byrne, W., Finke, M., Khudanpur, S., McDonnough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., 1998. Pronunciation modeling using a hand-labelled corpus for conversational speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 313–316.

Coleman, J., 1992. The phonetic interpretation of headed phonological structures containing overlapping constituents. Phonetics Yearbook 9, 1–44.

Crystal, D., 1995. The Cambridge Encyclopedia of the English Language. Cambridge University Press, Cambridge.

Dewey, G., 1923. Relative Frequency of English Speech Sounds. Harvard University Press, Cambridge, MA.

Doyle, A.C., 1892. The Adventures of Sherlock Holmes. Harper, New York.

Fosler, E., Weintraub, M., Wegmann, S., Kao, Y.-H., Khudanpur, S., Galles, C., Saraclar, M., 1996. Automatic learning of word pronunciation from data. In: Proceedings of the International Conference on Spoken Language Processing, pp. S28–29.

Fosler-Lussier, E., Morgan, N., 1998. Effects of speaking rate and word frequency on pronunciations in conversational speech. Speech Communication 29 (2–4), 137–158.

Fosler-Lussier, E., Greenberg, S., Morgan, N., 1999. Incorporating contextual phonetics into automatic speech recognition. In: Proceedings of the International Congress of Phonetic Sciences, San Francisco.

French, N.R., Carter, C.W., Koenig, W., 1930. The words and sounds of telephone conversations. Bell System Tech. J. 9, 290–324.

Ganapathiraju, A., Goel, V., Picone, J., Corrada, A., Doddington, G., Kirchhoff, K., Ordowski, M., Wheatley, B., 1997. Syllable – A promising recognition unit for LVCSR. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 207–214.

Gauvain, J., Lamel, L., Adda, G., Adda-Decker, M., 1994. The LIMSI continuous speech dictation system: Evaluation on the ARPA Wall Street Journal task. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 557–560.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 517–520.

Goldinger, S.D., Pisoni, D.B., Luce, P., 1996. Speech perception and spoken word recognition: Research and theory. In: Lass N. (Ed.), Principles of Experimental Phonetics, Mosby St. Louis, pp. 277–327.

Greenberg, S., 1997a. On the origins of speech intelligibility in the real world. In: Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels. Pont-a-Mousson, France, pp. 23–32.

Greenberg, S., 1997b. The switchboard transcription project. Research Report #24, Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series. Center for Language and Speech Processing. Johns Hopkins University Press, Baltimore, MD.

Greenberg, S., 1997c. Auditory function. In: Crocker, M. (Ed.), Encyclopedia of Acoustics. Wiley, New York, pp. 1301–1323.

Greenberg, S., 1998. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. In: Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Kerkrade, The Netherlands, pp. 47–56.

Greenberg, S., Hollenback, J., Ellis, D., 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In: Proceedings of the International Conference on Spoken Language Processing, Philadelphia, pp. S32–35.

Greenberg, S., Ellis, D.A., Hollenback, J., Fosler-Lussier, E., 1999. Phonetic transcription of spontaneous American English (the Switchboard corpus). Speech Communication (submitted).

Jespersen, O., 1922. Language; Its Nature, Development and Origin. Allen and Unwin, London.

Kahn, D., 1980. Syllable-based Generalizations in English Phonology. Garland, New York.

Kenyon, J.S., Knott, T.A., 1953. A Pronouncing Dictionary of American English. Merriam, Springfield, MA.

Kingsbury, B.E.D., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Communication 25, 117–132.

Kirchhoff, K., 1999. Robust speech recognition using articulatory information. Ph.D. thesis, University of Bielefeld.

Kohler, K., 1995. Articulatory reduction in different speaking styles. In: Proceedings of the International Congress of Phonetic Sciences, Stockholm, Vol. 2, pp. 12–19.

Kompe, R., 1997. Prosody in Speech Understanding Systems. Springer, Berlin.

Labov, W., 1972. Sociolinguistic Patterns. University of Pennsylvania Press, Philadelphia.

Lehiste, I., 1996. Suprasegmental features of speech. In: Lass, N. (Ed.), Principles of Experimental Phonetics. Mosby, St. Louis, pp. 226–244.

Levelt, W., 1989. Speaking. MIT Press, Cambridge, MA.

Lindblom, B., 1963. A spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773–1781.

Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H–H theory. In: Hardcastle, W., Marchal. A. (Eds.), Speech Production and Speech Modeling. Kluwer Academic Publishers, Dordrecht, pp. 403–439.

Lyovin, A., 1997. An Introduction to the Languages of the World. Oxford University Press, Oxford.

McAllaster, D., Gillick, L., Scattone, F., Newman, M., 1998. Explorations with fabricated data. In: Proceedings of the DARPA Workshop on Conversational Speech Recognition, Hub-5.

Niemann, H., Noth, E., Kiessling, A., Kompe, R., Batliner, A., 1997. Prosodic processing and its use in Verbmobil. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 75–78.

Ostendorf, M., Byrne, B., Macchiani, M., Finke, M., Gunawardana, A., Ross, K., Roweis, S., Shriberg, E., Talkin, D., Waibel, A., Wheatley, B., Zeppenfeld, T., 1997. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. Research Report #24, Large Vocabulary Continuous Speech Recognition Workshop Technical Report Series. Center for Language and Speech Processing. Johns Hopkins University, Baltimore, MD.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Riley, M., Ljolje, A., 1995. Automatic generation of detailed pronunciation lexicons. In: Lee, C.H., Soong, F.K., Paliwal, K.K. (Eds.), Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic Publishers, Boston.

Riley, M., Finke, M., Khudanpur, S., Llolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., 1998. Stochastic pronunciation modelling and hand-labelled phonetic corpora. In: Proceedings of the ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Kerkrade, pp. 109–116.

Schiel, F.A., Tillmann, H., 1998. Statistical modeling of pronunciation: it's not the model, it's the data. In: Proceedings of the ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Kerkrade, pp. 131–136.

Silipo, R., Greenberg, S., 1999. Automatic transcription of prosodic stress for spontaneous English discourse. In: Proceedings of the International Congress of Phonetic Sciences, San Francisco.

van Son, R.J.J.H., Koopmans-van Beinum, J., Pols, L.C.W., 1998. Efficiency as an organizing principle of natural speech. In: Proceedings of the International Conference on Spoken Language Processing, pp. 2375–2378.

van Kuik, D., Boves, L., 1999. Acoustic correlates of lexical stress in continuous telephone speech. Speech Communication 27, 95–111.

van Wieringen, A., 1995. Perceiving dynamic speechlike sounds. Ph.D. thesis, University of Amsterdam.

Waibel, A., 1988. Prosody and Speech Recognition. Morgan Kaufmann Publishers, San Mateo, CA.

Weintraub, M., Taussig, K., Smith, K.H., Snodgrass, A., 1996. Effect of speaking style on LVCSR performance. In: Proceedings of the International Conference on Spoken Language Processing, Philadelphia.

Weintraub, M., Fosler, E., Galles, C., Kao, Y.-H., Khudanpur, S., Saraclar, M., Wegmann, S., 1997. WS96 project report: Automatic learning of word pronunciation from data. Research Report #24, Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series. Center for Language and Speech Processing. Johns Hopkins University, Baltimore, MD.

Wu, S.-L., Kingsbury, B., Morgan, N., Greenberg, S., 1998a. Incorporating information from syllable-length time scales into automatic speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Seattle, pp. 721–724.

Wu, S.-L., Kingsbury, B., Morgan, N., Greenberg, S., 1998b. Performance improvements through combining phone- and syllable-length information in automatic speech recognition. In: Proceedings of the International Conference on Spoken Language Processing, Sydney, pp. 854–857.

Zipf, G.K., 1945. The meaning-frequency relationship of words. J. Gen. Psych. 33, 251–256.

Zue, V.W., Seneff, S., 1996. Transcription and alignment of the TIMIT database. In: Fujisaki, H. (Ed.), Recent Research Towards Advanced Man-Machine Interface Through Spoken Language. Elsevier, Amsterdam, pp. 515–525.